HOME^3:MULTI-ROOM GEOMETRY ESTIMATION VIA TOP-DOWN GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward 3D foundation models struggle in multi-room environments where cameras are segregated across rooms divided by opaque walls, breaking visual overlap and spatial continuity. We propose **HOME^3**, a method that adapts 3D foundation models to incorporate an unposed top-down RGB render alongside perspective captures for multi-room geometry and relative camera pose estimation. To support this, we construct **TD-HOMES**, a dataset suite spanning real-world scans, designer-crafted spaces, and agentic procedural environments. While providing global top-down context largely resolves catastrophic pose and geometry failures across disjoint rooms, naively sharing parameters induces representational interference that compromises pre-trained metric precision on standard overlapping scenes. To alleviate this issue, **HOME^3** introduces a view-type decoupling design that selectively decouples the projection weights and feed-forward networks within global cross-frame attention layers between top-down and perspective views. Experiments show that **HOME^3** reliably recovers coherent multi-room geometry across non-overlapping views while reducing performance drops on standard overlapping scenes.

1 INTRODUCTION

Estimating the relative spatial configuration between arbitrary camera viewpoints is a cornerstone problem in geometric computer vision. For decades, multi-view geometry pipelines, ranging from classical Structure-from-Motion systems (Schönberger & Frahm, 2016; Pan et al., 2024) to modern learning-based matchers (Sun et al., 2021; Edstedt et al., 2023; Leroy et al., 2024; He et al., 2025), have operated under the assumption that input views share sufficient visual overlap to establish local pixel correspondences. The recent emergence of feed-forward 3D foundation models (Wang et al., 2025a; 2026b; Lin et al., 2026) has shown promising potential, prompting efforts to fine-tune these architectures for extreme baselines with small to no overlap (Zhang et al., 2025; Li et al., 2026). However, extending them across multi-room indoor environments remains an open challenge (Fig. 1). In these settings, cameras are frequently segregated into separate rooms divided by opaque walls. Unlike outdoor or single-room extreme views where background continuity or shared structural planes can still guide alignment Cai et al. (2021); Bezalel et al. (2025), disjoint indoor spaces exhibit distinct architectural boundaries that break spatial continuity, leaving the inter-room relative pose unconstrained and susceptible to large configuration ambiguities.

To connect these disjoint viewpoints, the network requires an external cue, i.e., an exocentric structural prior that captures the macroscopic spatial configuration of the environment. While such global priors can take multiple forms, such as 2D architectural blueprints or coarse geometric scaffolds, we instantiate this prior as a top-down RGB render. Crucially, we treat this render strictly as an unposed 2D visual guide without known camera poses or metric scaling relative to the perspective observations. This reflects a considerably more practical setting, as real-world architectural sketches or top-down renders rarely provide precise spatial alignments Huang et al. (2025). Although existing indoor datasets provide rich perspective observations, they lack top-down (TD) context capturing the global spatial configuration.

Consequently, we construct **TD-HOMES**, an indoor dataset spanning three distinct scene paradigms: real-world scans from HM3D (Ramakrishnan et al., 2021), designer-crafted synthetic spaces from Structured3D (Zheng et al., 2020), and agentic procedural environments generated via

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

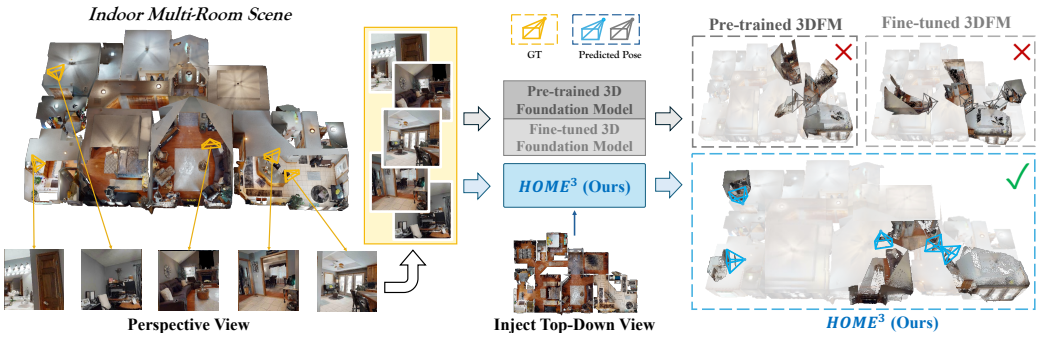

Figure 1: **Multi-room indoor scene reconstruction from disjoint captures.** When perspective images are sampled across disconnected rooms in a multi-room environment, standard 3D foundation models, both off-the-shelf and fine-tuned, struggle to resolve cross-room spatial relationships relying solely on perspective observations, leading to collapsed camera poses and distorted geometry. In contrast, **HOME³** leverages an unposed top-down render as an exocentric spatial guide, bridging non-overlapping views to recover globally consistent camera poses and dense geometry.

Code-as-Room (Yang et al., 2026). To pair perspective captures with top-down views across different data sources, we develop tailored extraction pipelines based on their geometric characteristics. However, leveraging this top-down prior is non-trivial: both traditional pixel-to-pixel matches and the implicit feature associations learned by feed-forward models constantly fail to connect these two types of views (see Fig. 2 and Appendix A.1).

We therefore propose **HOME³**, a method that adapts 3D foundation models to incorporate top-down renders into multi-room geometry estimation. We first ingest the top-down render alongside perspective images and jointly fine-tune the model end-to-end using TD-HOMES, and show that providing global top-down context largely resolves catastrophic camera pose and geometry regression failures across disjoint rooms. However, forcing such disparate observations through shared attention projection matrices and feed-forward layers induces parameter interference. This conflict not only hinders the network from fully leveraging top-down spatial cues for disjoint-view reasoning, but also compromises its pre-trained metric precision on standard overlapping scenes.

To overcome this representational conflict, we further introduce *view-type decoupling* for better multi-room geometry estimation. Instead of passing divergent viewpoints through shared weights, we disentangle in-situ perspective and top-down tokens using view-specific projection matrices (W_Q, W_K, W_V) and specialized feed-forward networks within each layer, while maintaining cross-view self-attention interaction. We primarily validate this design using π^3 (Wang et al., 2026b), while also verifying that it can be transferred to other 3D foundation models, e.g., VGGT-Ω (Wang et al., 2026a). Through extensive experiments, we show that our method effectively leverages visual top-down information to infer multi-room geometry among visually disjoint, non-overlapping views, whilst alleviating degradation of the original capabilities on standard overlapping scenes.

2 Related Work

2.1 Feed-Forward Geometry Prediction

Recent feed-forward approaches predict dense geometry and camera poses directly from unposed images, spanning pairwise reconstructors (Wang et al., 2024; Yang et al., 2025), streaming frameworks (Wang et al., 2025b; Lan et al., 2026), and generalist 3D foundation models (Wang et al., 2025a; Liu et al., 2025; Wang et al., 2026b; Lin et al., 2026; Wang et al., 2026a). As analyzed by Bratulić et al. (2026), these models capture epipolar geometry and cross-view correspondences through alternating attention, echoing classical and learning-based matching (Sun et al., 2021; Edstedt et al., 2023; He et al., 2025; Lowe, 2004). To complement RGB inputs, subsequent methods incorporate auxiliary signals such as view-wise depth maps and sensor priors (Jang et al., 2025; Liu et al., 2025; Peng et al., 2025; Keetha et al., 2026). Different from them, we treat the input top-down view as an unaligned, exocentric observation that serves as a global spatial bridge across non-overlapping, multi-room perspective views.

2.2 Extreme Pose Estimation

Extreme pose estimation is challenging due to limited visual overlap (Cai et al., 2021; Dekel et al., 2024), extends them to in-the-wild imagery (Bezalel et al., 2025), and exploits diffusion models to capture 3D relationships (Wang et al., 2023; Zhang et al., 2024). Recent studies reveal emergent cross-view reasoning in 3DFMs and improve relative rotation estimation through lightweight fine-tuning (Zhang et al., 2025), but also identify key failure modes under sparse, unconstrained captures (Li et al., 2026). Despite these advances, perspective-only inputs and standard fine-tuning remain insufficient for robust estimation across non-overlapping views. We introduce top-down RGB renders as an exocentric guide to bridge disjoint views during multi-room reconstruction.

2.3 Floorplan Localization

Floorplan localization establishes spatial correspondences between the perspective observations and the structural scene layout of the whole scene. Prior methods (Howard-Jenkins et al., 2021; Howard-Jenkins & Prisacariu, 2022) compact the floorplan and perspective captures into a shared latent space and register the localization by comparing the features. Several methods (Chen et al., 2024; Wüest et al., 2026) implement learned observation models, combined with sequential filtering to integrate localization evidence over time. Huang et al. (2025) show that existing feed-forward geometric models (Wang et al., 2024) can establish point-level correspondences between images and floorplans, framing localization as a geometric registration problem. These studies suggest that global scene layouts provide complementary spatial cues beyond those available from perspective observations alone. In our work, we leverage a top-down render as an exocentric guide for the 3D foundation model, enabling geometric estimation across complex multi-room scenes.

3 TD-Homes Dataset

Existing indoor datasets primarily provide in-situ perspective observations, but lack scene-level top-down context that captures the global spatial configuration of the environment. To provide such complementary structural priors, we introduce **TD-HOMES**, a diverse indoor dataset suite spanning three distinct scene paradigms: real-world scans, designer-crafted synthetic spaces, and agentic procedural environments. An overview is provided in Fig. 2. Specifically, we construct floor-wise top-down visual priors alongside perspective captures across these domains, building upon sources such as HM3D (Ramakrishnan et al., 2021), Structured3D (Zheng et al., 2020), and procedural generation via the Code-as-Room (Yang et al., 2026) framework.

Real-World Scans. Real-world multi-storey scans in HM3D (Ramakrishnan et al., 2021) span continuous vertical levels without explicit floor-level decomposition or top-down layout imagery. To generate floor-aligned top-down priors, we first partition each multi-storey capture into independent, single-floor subscenes by analyzing camera elevation histograms and vertical bounding boundaries from the reconstructed 3D mesh. We assign in-situ RGB images, depth maps, and camera poses to their respective floor subscenes. For each floor, the 3D mesh is sliced at the ceiling level to eliminate roof occlusions, aligned to standard camera coordinates, and rendered from an overhead orthographic viewpoint looking directly down to produce an uncalibrated layout map.

Designer-Crafted Synthetic Scenes. While Structured3D (Zheng et al., 2020) originates from professionally designed CAD environments, the released benchmark distributes rendered panoramas, perspective views, and analytical geometric primitives rather than complete 3D scene meshes. Consequently, we cannot directly render overhead views from an existing mesh. Instead, we reconstruct the top-down layout priors via a structured projection pipeline: we first parse the provided line and junction annotations to identify wall and floor planes, preserve structural openings (doors and windows), and assemble closed 3D room boundary polygons. We then inverse-project the high-resolution panoramic RGB textures onto these reconstructed room polygons. Finally, we capture the top-down prior using an elevated weak-perspective camera combined with height-based ceiling slicing and horizontal coordinate alignment, yielding clean, unoccluded top-down layout imagery paired with the original perspective captures.

Agentic Procedural Environments. We construct 3D environments by combining floor-level layout images from the MANSION dataset (Che et al., 2026) with the agentic synthesis pipeline of

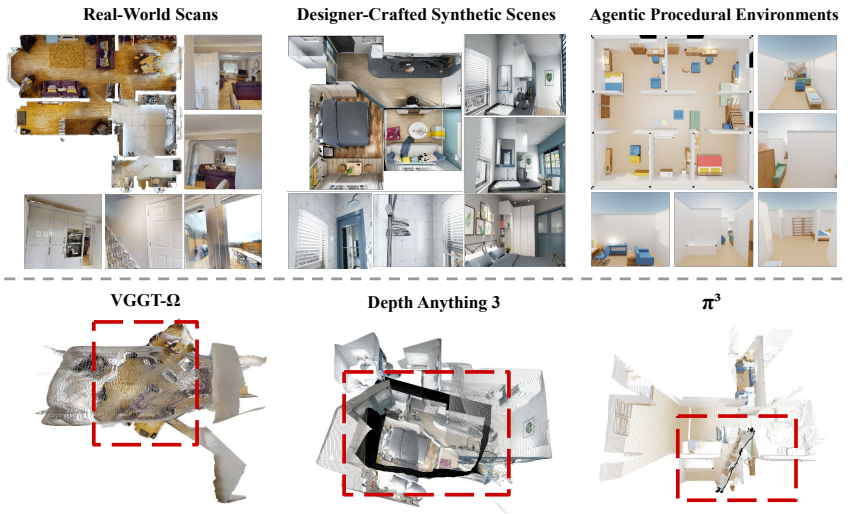

Figure 2: **Overview of the TD-HOMES dataset. Top:** Example scenes across three distinct domains: real-world scans (Ramakrishnan et al., 2021), designer-crafted spaces (Zheng et al., 2020), and agentic procedural environments (Yang et al., 2026)), displaying a central top-down view surrounded by sampled in-situ perspective captures. These top-down views are generated by slicing ceiling geometry to eliminate roof occlusions and rendering an orthographic projection looking directly down. **Bottom:** Off-the-shelf 3D foundation models (Wang et al., 2026a; Lin et al., 2026; Wang et al., 2026b) struggle when fed such top-down views alongside perspective images, failing to resolve multi-room geometry and relative camera poses.

Code-as-Room (Yang et al., 2026). We first curate top-down floor-level maps from MANSION, prioritizing scenes with multi-room layouts and complex spatial topologies while filtering out simple single-room configurations. For each selected top-down image, we execute the coarse-to-fine agentic MLLM workflow from Code-as-Room: the agent parses functional zones, structural boundaries, object hierarchies, and spatial relationships into an executable layout program, which is progressively populated with object geometry, materials, textures, and lighting in Blender. Finally, we extract the top-down layout render from each generated 3D room to pair with the scene's perspective captures.

4 METHOD

In this section, we present our framework for estimating relative camera poses in complex indoor environments, with a particular focus on resolving visually disconnected scenes where co-visibility collapses. To bridge these disjoint viewpoints, we incorporate a scene-level top-down render to provide exocentric structural cues. We first review the feed-forward geometric model baseline and analyze its failure modes under multi-room scenarios in Sec. 4.1. Sec. 4.2 formalizes our method **HOME³**, top-down guided formulation, examines the representational trade-offs of naive joint fine-tuning. Then introduces view-type decoupling, which is an architectural design that separates projection matrices and feed-forward layers across divergent view types to incorporate top-down context while helping to alleviate performance degradation on standard overlapping scenes.

4.1 PRELIMINARIES

Feed-forward geometric models (Wang et al., 2026a;b; 2025a) estimate camera poses and dense scene geometry in a direct forward pass. Given a set of N perspective views $\mathcal{S} = \{I_i\}_{i=1}^{N}$, a pre-trained visual backbone first extracts 2D patch tokens. These representations are then processed through Transformer blocks that alternate between intra-frame self-attention and cross-frame global attention to correlate visual features across viewpoints. Within each cross-frame block, tokens undergo multi-head self-attention followed by feed-forward networks (FFNs). Finally, the network regresses the 3D scene geometry $\hat{\mathcal{G}} = \phi(\mathcal{S})$, yielding relative camera poses $\hat{T}_i \in \mathrm{SE}(3)$ alongside dense pixel-aligned pointmaps $\hat{X}_i \in \mathbb{R}^{H\times W\times 3}$ for each view I_i.

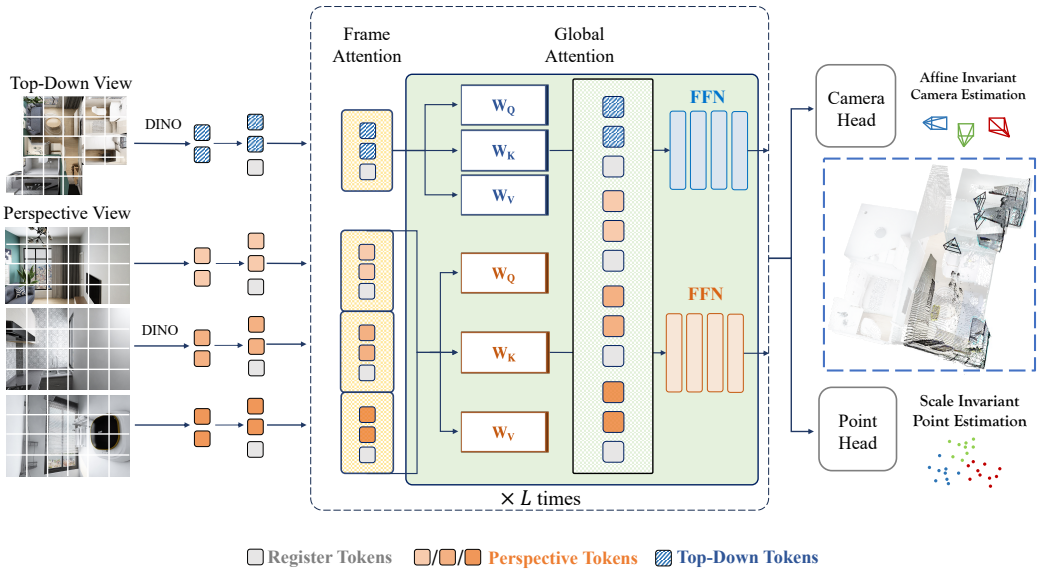

Figure 3: **Illustration of View-Type Decoupling of HOME^3.** Given an unposed top-down render and multiple in-situ perspective images, each view is patchified and encoded using a shared DINO backbone with appended register tokens. The token sequences are processed through L alternating blocks of intra-frame and cross-frame attention. Within each intra-frame layer, self-attention operates locally per view. In the subsequent cross-frame layer, tokens from the top-down render and perspective images are projected using separate linear weights (W_Q, W_K, W_V) before computing full cross-attention across all tokens, followed by decoupled feed-forward networks (FFNs) for each view type. Finally, the updated perspective tokens are decoded by dedicated prediction heads to regress affine-invariant camera poses and scale-invariant 3D pointmaps.

While remarkably effective across scenes with sufficient visual overlap, this feed-forward paradigm struggles under multi-room configurations. When views are distributed across non-overlapping rooms, the absence of visual co-visibility deprives the cross-frame attention layers of the photometric correspondences and continuous epipolar constraints they rely on. As illustrated in Fig. 1, this leads to catastrophic scale and pose drift across disjoint spaces. Furthermore, fine-tuning existing architectures exclusively on perspective multi-room sequences cannot fundamentally resolve this ambiguity, as evidenced by the failure cases in Fig. 1; no amount of optimization on perspective-only data can hallucinate the macroscopic spatial layout absent from the input observations.

4.2 HOME^3

To resolve the unconstrained spatial ambiguity across disjoint rooms, the network requires an exocentric guidance $\mathcal{E}$ that captures macroscopic room topology. However, existing feed-forward geometry models struggle to exploit top-down guidance during zero-shot inference, as demonstrated in Fig. 2. We therefore propose **HOME^3**, a novel method that adapts 3D foundation models for top-down view guided multi-room geometry prediction. Formally, we adapt the model from directly learning $\hat{\mathcal{G}} = \phi(\mathcal{S})$ to $\hat{\mathcal{G}} = \{\hat{\mathbf{T}}, \hat{\mathbf{X}}\} = \psi(\mathcal{S}, \mathcal{E})$, where ψ denotes our geometry predictor, and $\hat{\mathbf{T}}$ and $\hat{\mathbf{X}}$ denote the predicted camera poses and dense pointmaps for $\mathcal{S}$, respectively. We instantiate $\mathcal{E}$ as a top-down RGB render that captures global room topology, used solely as a visual guide without requiring its camera parameters or metric alignment with the perspective observations. To motivate our adaptation design, we first examine joint fine-tuning with shared parameters and its limitations before introducing view-type decoupling.

Joint Fine-Tuning. Jointly fine-tuning a pretrained 3DFM on perspective observations and the top-down render offers a viable approach. Specifically, perspective tokens $\mathbf{H}_{\mathcal{S}}$ and top-down tokens $\mathbf{H}_{\mathcal{E}}$ are concatenated into $[\mathbf{H}_{\mathcal{S}}\,;\,\mathbf{H}_{\mathcal{E}}]$ and jointly processed in the global Transformer blocks using shared projection matrices and FFNs. This baseline substantially alleviates cross-room camera pose failures, highlighting the value of top-down context for constraining inter-room configurations.

However, these gains come at the cost of reduced pretrained geometric precision on standard overlapping scenes (Tab. 3). We conjecture that this degradation arises from parameter interference when shared projection matrices and FFNs must accommodate the distinct characteristics of perspective and top-down observations. In particular, matches between top-down and perspective views are difficult to establish: existing advanced feature matchers (Sun et al., 2021; Edstedt et al., 2023; He et al., 2025) often fail across these viewpoints (see Appendix A.1). Lacking reliable visual correspondences to anchor the top-down views, sharing weights likely exerts conflicting gradients on the network, which may disrupt the pre-trained geometric representations.

View-Type Decouple. To address this trade-off, **HOME^3** retains the local attention layers while replacing the shared QKV projections and FFNs with view-type-specific counterparts within global attention layers. All remaining model parameters stay shared between input perspective and top-down views. This design accommodates the distinct characteristics of top-down and perspective observations through view-specific parameters while preserving interaction through global attention.

Specifically, at the input of each cross-frame attention layer, the perspective tokens $\mathbf{H}_{\mathcal{S}}$ and top-down tokens $\mathbf{H}_{\mathcal{E}}$ are projected into queries, keys, and values using separate parameter matrices:

$$\begin{aligned} \mathbf{Q}_{\mathcal{S}}, \mathbf{K}_{\mathcal{S}}, \mathbf{V}_{\mathcal{S}} &= \mathbf{H}_{\mathcal{S}} W_Q^{\mathcal{S}}, \mathbf{H}_{\mathcal{S}} W_K^{\mathcal{S}}, \mathbf{H}_{\mathcal{S}} W_V^{\mathcal{S}}, \\ \mathbf{Q}_{\mathcal{E}}, \mathbf{K}_{\mathcal{E}}, \mathbf{V}_{\mathcal{E}} &= \mathbf{H}_{\mathcal{E}} W_Q^{\mathcal{E}}, \mathbf{H}_{\mathcal{E}} W_K^{\mathcal{E}}, \mathbf{H}_{\mathcal{E}} W_V^{\mathcal{E}}. \end{aligned} \tag{1}$$

Because the two view types use separate parameter sets for these transformations, their parameters can specialize to the respective viewing characteristics while remaining coupled through shared global attention interactions.The projected queries, keys, and values are then concatenated along the sequence dimension and processed via a unified global attention operation:

$$\mathbf{Z} = \text{Softmax}\left(\frac{[\mathbf{Q}_{\mathcal{S}}\,;\,\mathbf{Q}_{\mathcal{E}}][\mathbf{K}_{\mathcal{S}}\,;\,\mathbf{K}_{\mathcal{E}}]^T}{\sqrt{d}}\right)[\mathbf{V}_{\mathcal{S}}\,;\,\mathbf{V}_{\mathcal{E}}]\,W_O, \tag{2}$$

where W_O is the shared output projection matrix.

Following attention aggregation, the unified representations $\mathbf{Z}$ are partitioned back into perspective features $\mathbf{Z}_{\mathcal{S}}$ and top-down features $\mathbf{Z}_{\mathcal{E}}$ according to their original sequence lengths, and subsequently passed through independent feed-forward networks with residual connections. This design allows top-down contextual cues to guide perspective geometry estimation while alleviating parameter interference between the two view types.

Training Objectives. Following standard practice, we apply camera pose and pointmap regression losses exclusively to the perspective views $\mathcal{S}$. Predictions corresponding to the top-down tokens are masked out from the loss, allowing the top-down render to act purely as a spatial guide without requiring ground-truth 3D supervision. While ground-truth rotation is available for the top-down view, we find that supervising it yields no empirical benefit and can degrade pose estimation across perspective views (see Sec. 5.3).

5 Experiments

5.1 Experimental Setup

Dataset Preparation. We train and evaluate on TD-HOMES, spanning three complementary sources: Real-World Scans, Designer-Crafted Synthetic Scenes, and Agentic Procedural Environments. Across the real-world and synthetic collections, we source 1.6K and 3K floor-level scenes comprising roughly 8.5K and 18K rooms, containing 1.7M and 168K perspective views, respectively. Camera poses span broadly across each environment, naturally providing view pairs with both overlapping and disjoint coverage. We train on both subsets using a 9:1 split, reserving held-out scenes for in-domain evaluation. The agentic procedural environments provide 80 unseen residential scenes, each paired with a top-down render and 49 perspective images (3,920 views in total), reserved exclusively for zero-shot testing. To evaluate whether introducing top-down renders affects standard multi-view reconstruction, we further benchmark on ETH3D (Schops et al., 2017), RealEstate10K (Zhou et al., 2018), and Neural RGB-D (Azinović et al., 2022) under standard perspective-only settings.

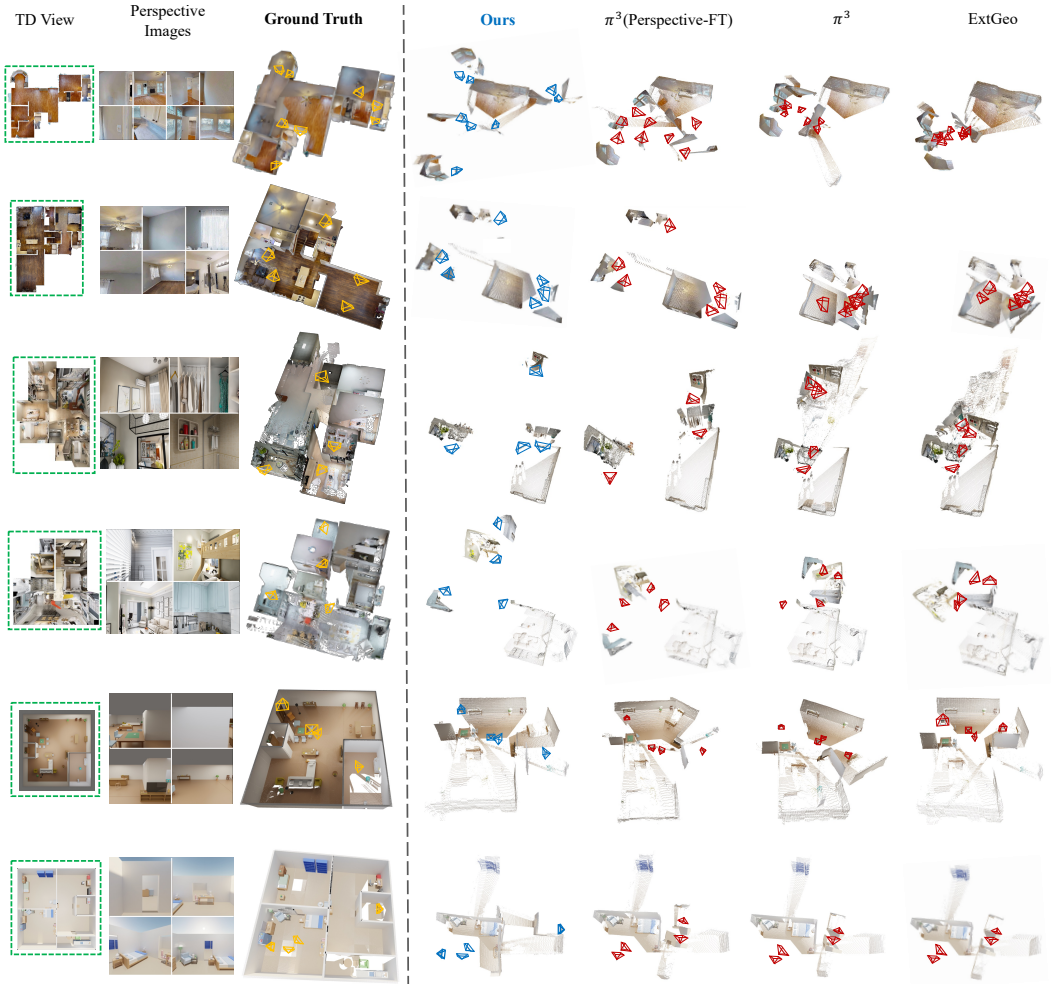

Figure 4: **Qualitative comparison of geometry estimation on multi-room scenes.** We visualize both predicted camera poses (blue/red cameras) and 3D pointmaps against the ground truth (yellow cameras). Existing pre-trained and fine-tuned baselines struggle to establish cross-room spatial relationships. In contrast, our **HOME^3** w/ *view-type decoupling* leverages the unposed top-down render to reliably estimate spatial relationships across non-overlapping perspective views, recovering more coherent global scene geometry.

Training Implementation. We build our method **HOME^3** upon π^3 (Wang et al., 2026b), which serves as the primary backbone for all main experiments, baseline variants, and ablation studies. To verify that our findings generalize across architectures, we additionally extend both joint fine-tuning and view-type decoupling to VGGT-Ω (Wang et al., 2026a). In all cases, networks are initialized from their respective pre-trained weights. When applying view-type decoupling, the separate top-down projection matrices and feed-forward layers are initialized as copies of their corresponding perspective counterparts. Specific data mix ratio, training batch construction protocols and optimization hyperparameters are detailed in Appendix A.2.

Comparison Methods. We evaluate our approach against several representative configurations to isolate the contributions of in-domain fine-tuning, extreme-view tuning, and view-type decoupling. On the primary π^3 (Wang et al., 2026b) backbone, we compare against: (1) **Off-the-shelf π^3**, evaluating zero-shot inference without modification; (2) **ExtGeo** (Zhang et al., 2025), a recent extreme-view adaptation baseline that tunes π^3 to enhance non-overlapping pose estimation; and (3) **Perspective-only fine-tuning**, which trains the backbone exclusively on our perspective training views without top-down renders, isolating the benefit of the data corpus itself. Furthermore, to

evaluate architectural generality, we compare our adapted **VGGT-Ω** (Wang et al., 2026a) (equipped with joint fine-tuning and view-type decoupling) against its original pre-trained baseline.

5.2 Multi-Room Geometry Estimation

We evaluate all methods on dense point map reconstruction and relative camera pose estimation across both in-domain and zero-shot environments. Quantitative results are reported in Tab. 1 and Tab. 2, with visual comparisons shown in Fig. 4; additional benchmark results and visual analyses are provided in Appendix A.3.

Table 1: **Point Map Estimation on TD-HOMES.** TD-HOMES (zero-shot) denotes the unseen agentic procedural environments. Acc., Comp., and N.C. denote point map accuracy, completeness, and normal consistency, respectively. **HOME^3** in blue and **HOME^3** in green use π^3 and VGGT-Ω as backbones, respectively.

Method	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
ExtGeo (Zhang et al., 2025)	0.653	0.601	1.179	1.022	0.607	0.659	0.296	0.262	0.441	0.372	0.749	0.846
π^3 (Wang et al., 2026b)	0.653	0.599	1.206	1.049	0.612	0.665	0.281	0.247	0.436	0.369	0.766	0.855
π^3 (Perspective-FT)	0.567	0.497	0.805	0.661	0.662	0.749	0.289	0.236	0.430	0.349	0.735	0.852
HOME^3 (Joint-FT)	0.295	0.240	0.362	0.277	0.765	0.889	0.194	0.156	0.218	0.173	0.801	0.928
HOME^3 (Decouple)	**0.278**	**0.223**	**0.333**	**0.252**	**0.773**	**0.897**	**0.161**	**0.132**	**0.178**	**0.148**	**0.826**	**0.948**
VGGT-Ω (Wang et al., 2026a)	0.639	0.587	1.173	1.021	0.615	0.665	0.277	0.247	0.476	0.416	0.767	0.843
HOME^3 (Joint-FT)	0.318	0.259	0.400	0.307	0.746	0.864	0.163	0.135	0.172	0.139	0.837	0.950
HOME^3 (Decouple)	**0.302**	**0.246**	**0.367**	**0.277**	**0.752**	**0.873**	**0.152**	**0.122**	**0.154**	**0.117**	**0.840**	**0.958**

Point Map Estimation. In line with prior works (Wang et al., 2026b; Zhang et al., 2025), we report Accuracy (Acc.), Completion (Comp.), and Normal Consistency (N.C.) to assess the performance of point map estimation. Tab. 1 indicates that **HOME^3** surpasses previous methods. The advantage remains pronounced under zero-shot evaluation. Qualitatively, our **HOME^3** produces coherent geometry that matches the ground truth more faithfully (Fig. 4). These results demonstrate that our design enhances accuracy and robustness, generalizing to unseen indoor environments.

Relative Camera Pose Estimation. Following previous works (Wang et al., 2026b; Zhang et al., 2025), we evaluate Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) at certain thresholds, and Area Under the Curve (AUC). Fig. 4 shows that **HOME^3** accurately recovers camera configurations from spatially disjoint images where previous SOTA methods fail (Fig. 2). Tab. 2 shows that, although π^3 (Wang et al., 2026b) surpasses the pre-trained model, it remains insufficient for multi-room geometry estimation. *ExtGeo* performs worse than vanilla π^3 due to its minimal fine-tuning strategy and domain shift. Both of them degrade under zero-shot scenes. **HOME^3** consistently outperforms previous methods and further shows that view-type decoupling benefits joint geometry prediction.

Generalization Analysis. Beyond its effectiveness on π^3 (Tabs. 1 and 2), **HOME^3** also yields favorable results when adapted to other 3D foundation models, such as VGGT-Ω. Qualitative results of adapting to VGGT-Ω can be found in Appendix. A.3.1. Crucially, the relative behavior of view-type decoupling compared to joint fine-tuning shows a consistent tendency across both backbones. Similar to observations on π^3, applying view-type decoupling to VGGT-Ω helps reduce performance drops on inputs without top-down renders, outperforming Joint-FT across most benchmarks in perspective-only settings (Tab. 3). This consistent pattern suggests that the benefits of decoupling view-specific parameters are not restricted to a single architecture.

5.3 Ablation Study

Ablation of View-Type Decoupling. We ablate view-type decoupling across local and global attention layers. As reported in Appendix A.3.2, global-only decoupling yields the best overall point

Table 2: **Camera Pose Estimation on TD-HOMES.** RRA, RTA, and AUC denote Relative Rotation Accuracy, Relative Translation Accuracy, and Area Under the Curve, respectively, evaluated at a 30° threshold. **HOME^3** in blue and **HOME^3** in green use π^3 and VGGT-Ω as backbones, respectively.

Method	TD-HOMES			TD-HOMES (zero-shot)		
	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑	RRA@30 ↑	RTA@30 ↑	AUC@30 ↑
ExtGeo (Zhang et al., 2025)	54.14	54.77	30.19	96.03	83.86	76.91
π^3 (Wang et al., 2026b)	56.06	55.99	32.09	96.91	84.56	79.35
π^3 (Perspective-FT)	69.75	68.09	44.65	97.32	84.65	74.17
HOME^3 (Joint-FT)	87.57	87.06	69.14	98.20	89.25	80.06
HOME^3 (Decouple)	**88.70**	**87.96**	**71.46**	**99.05**	**89.83**	**83.28**
VGGT-Ω	54.52	55.49	30.85	97.93	83.34	78.23
HOME^3 (Joint-FT)	85.64	85.21	65.64	98.78	**89.70**	**84.83**
HOME^3 (Decouple)	**86.95**	**86.16**	**67.76**	**99.05**	**89.70**	83.16

Table 3: **Decoupled adaptation preserves pre-trained perspective capabilities.** All methods are evaluated on perspective-only inputs across ETH3D, RealEstate10K, and sparse NRGBD sequences. RRA, RTA, and AUC are evaluated at a 15° threshold. **HOME^3** in blue and **HOME^3** in green use π^3 and VGGT-Ω as backbones, respectively.

Method	ETH3D			Re10K			NRGBD		
	RRA@15 ↑	RTA@15 ↑	AUC@15 ↑	RRA@15 ↑	RTA@15 ↑	AUC@15 ↑	RRA@15 ↑	RTA@15 ↑	AUC@15 ↑
π^3 (Wang et al., 2026b)	99.99	78.02	61.65	99.84	92.59	80.60	99.99	99.99	86.19
HOME^3 (Joint-FT)	99.99	78.85	59.23	99.84	92.15	78.08	99.99	92.86	83.10
HOME^3 (Decouple)	99.99	82.97	61.10	99.84	92.31	78.42	99.99	96.43	86.19
VGGT-Ω (Wang et al., 2026a)	99.99	97.88	84.83	99.84	93.06	80.57	99.99	99.99	90.71
HOME^3 (Joint-FT)	99.99	96.56	78.62	99.84	91.99	77.43	99.99	96.43	86.90
HOME^3 (Decouple)	99.99	96.83	82.06	99.84	92.12	78.23	99.99	99.99	87.38

map reconstruction in both in-domain and zero-shot settings. Decoupling both layer types slightly improves in-domain pose estimation, whereas our design generalizes better, increasing zero-shot AUC@5° ↑ from 52.54 to 59.68.

Ablation of Top-Down Rotation Supervision. We investigate whether the top-down render should receive direct geometric supervision during training. Canonical top-down alignment makes these rotation targets available, whereas metrically aligned translation and depth targets are unavailable. As shown in Appendix A.3.2, training without the top-down rotation supervision improves overall relative pose estimation and point map reconstruction. Under zero-shot evaluation, omitting the top-down rotation supervision reduces mean rotation error from 3.08° to 2.37°.

6 Conclusion

In this work, we tackled feed-forward geometry and camera pose estimation across multi-room indoor environments. We introduced **HOME^3**, adapting 3D foundation models to leverage unposed top-down RGB renders to bridge disconnected rooms separated by opaque walls. To facilitate training and evaluation, we constructed TD-HOMES across real, synthetic, and procedural indoor spaces. Furthermore, to mitigate parameter interference from processing disparate view types, we presented a view-type decoupling design that assigns specialized projection weights and feed-forward networks in global cross-frame layers while preserving cross-view interaction. Extensive experiments demonstrate that **HOME^3** reliably recovers coherent multi-room geometry and camera poses across visually disjoint viewpoints while reducing performance drops on standard perspective captures. In the future, we plan to incorporate more diverse external guidance modalities, such as 2D architectural blueprints and hand-drawn layout sketches, generalize to unconstrained in-the-wild environments, and validate our decoupling framework across a broader range of 3D foundation models.

7 AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets and polishing writing. We have not used generative AI tools for [retrieval and discovery, research ideation or execution, draft sections of the paper], and [proving mathematical claims] is not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

8 REPRODUCIBILITY STATEMENT

We have taken several steps to facilitate the reproducibility of our work. The description of the proposed architecture and training objectives can be found in Sec. 4. The construction and preprocessing procedures of the datasets are provided in Sec. 3. The training and evaluation protocols of the used datasets are documented in Sec. 5 and Appendix A.2.1. Detailed implementation settings are reported in Appendix A.2, including hyperparameters and computational resources. The source code, data processing scripts, and evaluation code will be made publicly available upon acceptance to facilitate reproduction of the reported results.

REFERENCES

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbach-Elor. Extreme rotation estimation in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jelena Bratulić, Sudhanshu Mittal, Thomas Brox, and Christian Rupprecht. On geometric understanding and learned priors in feed-forward 3d reconstruction models. In *European Conference on Computer Vision*, 2026.

Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Lirong Che, Shuo Wen, Shan Huang, Chuang Wang, Yuzhe Yang, Gregory Dudek, Xueqian Wang, and Jian Su. Mansion: Multi-floor language-to-3d scene generation for long-horizon tasks. *arXiv preprint arXiv:2603.11554*, 2026.

Changan Chen, Rui Wang, Christoph Vogel, and Marc Pollefeys. F3loc: Fusion and filtering for floorplan localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xingyi He He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. *ArXiv*, 2025.

Henry Howard-Jenkins and Victor Adrian Prisacariu. LaLaLoc++: Global floor plan comprehension for layout localisation in unvisited environments. In *The European Conference on Computer Vision*, 2022.

Henry Howard-Jenkins, Jose-Raul Ruiz-Sarmiento, and Victor Adrian Prisacariu. Lalaloc: Latent layout localisation in dynamic, unvisited environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Kuan Wei Huang, Brandon Li, Bharath Hariharan, and Noah Snavely. C3po: Cross-view cross-modality correspondence by pointmap prediction. In *Advances in Neural Information Processing Systems*, 2025.

Wonbong Jang, Philippe Weinzaepfel, Vincent Leroy, Lourdes Agapito, and Jerome Revaud. Pow3r: Empowering unconstrained 3d reconstruction with camera and scene priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. In *International Conference on 3D Vision*, 2026.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. STream3R: Scalable sequential 3D reconstruction with causal transformer. In *ICLR*, 2026.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, 2024.

Yuan Li, Yuanbo Xiangli, Hadar Averbuch-Elor, Noah Snavely, and Ruojin Cai. Long-tail internet photo reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations*, 2026.

Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting. *arXiv preprint arXiv:2510.10726*, 2025.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004.

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L. Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, 2024.

Haosong Peng, Hao Li, Yalun Dai, Yushi Lan, Yihang Luo, Tianyu Qi, Zhengshen Zhang, Yufeng Zhan, Junfei Zhang, Wenchao Xu, and Ziwei Liu. Omnivggt: Omni-modality driven visual geometry grounded transformer. *arXiv preprint arXiv:2511.10560*, 2025.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-■. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. \piˆ3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026b.

Matthias Wüest, Francis Engelmann, Ondrej Miksik, Marc Pollefeys, and Daniel Barath. Unloc: Leveraging depth uncertainties for floorplan localization. In *International Conference on Learning Representations*, 2026.

Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yixuan Yang, Zhen Luo, Wanshui Gan, Jinkun Hao, Junru Lu, Jinghao Yan, Zhaoyang Lyu, and Xudong Xu. Code-as-room: Generating 3d rooms from top-down view images via agentic code synthesis. *arXiv preprint arXiv:2605.18451*, 2026.

Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations*, 2024.

Yiwen Zhang, Joseph Tung, Ruojin Cai, David Fouhey, and Hadar Averbuch-Elor. Emergent extreme-view geometry in 3d foundation models. *arXiv preprint arXiv:2511.22686*, 2025.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, 2020.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 2018.

A Appendix

A.1 Feature Matching Between Top-Down and Perspective Views

Incorporating top-down renders alongside perspective captures is non-trivial, as conventional correspondence-based pipelines fail to bridge these two view types. As shown in Fig. 5, advanced feature matchers (Sun et al., 2021; Edstedt et al., 2023; He et al., 2025) struggle to find correct pixel-to-pixel matches between the top-down render and in-situ perspective images. The radical change in viewing angle, combined with minimal visible surface overlap and differing projective characteristics, prevents standard matching algorithms from establishing reliable geometric correspondences across rooms. Furthermore, existing feed-forward 3D foundation models fail when naively fed top-down views alongside perspective images, collapsing relative camera poses and producing distorted geometry. These failures highlight that implicit feature associations learned by off-the-shelf architectures are insufficient to connect such heterogeneous viewpoints.

A.2 Implementation Details

A.2.1 Viewpoint Sampling Details

Three Complementary Overlap Levels. High-overlap sequences capture smooth, continuous trajectories within localized areas where adjacent frames maintain dense visual co-visibility, enabling the network to learn fine-grained geometric correspondences. Arbitrary-overlap sequences follow the sampling conventions of geometric foundation models (Wang et al., 2026b), reflecting unconstrained camera walks obtained via standard temporal strides or neighborhood indexing. Finally, Zero-overlap sequences enforce global spatial isolation, where every sampled view shares virtually no co-visibility with any other frame in the set.

Algorithm 1 Large-overlapping and Non-overlapping Pair Selection

Input:
Distance between views (r)
Viewing angle (θ)
Solution vector ($\mathbf{s} = [s_1, s_2]^\top$)
Diagonal FoV (Θ_{dia}), Horizontal FoV (Θ_h)
Camera frustums ($\mathcal{F}_i, \mathcal{F}_j$) with Far Depth ($d_{\mathcal{F}}$)
Output: Co-visibility type: High-overlap, or Zero-overlap

```
if r < r_pos or (r < r_Fpos and F_i ∩ F_j ≠ ∅) then
    if s_1 > 0 and s_2 > 0 and θ < θ_front then
        return High-overlap
    if s_1 < 0 and s_2 < 0 and θ < Θ_h then
        return High-overlap
    if s_1 × s_2 ≤ 0 and θ < Θ_h then
        return High-overlap
if r > r_neg then
    return Zero-overlap (Far distance)
else
    if s_1 > 0 and s_2 > 0 and θ > θ_opp then
        return Zero-overlap (Facing opposite directions)
    else if s_1 < 0 and s_2 < 0 and θ > Θ_dia then
        return Zero-overlap (Diverging views)
    else if s_1 × s_2 ≤ 0 and F_i ∩ F_j = ∅ then
        return Zero-overlap (Disjoint frustums)
```

To instantiate these sequences, we classify candidate frame pairs (I_i, I_j) using their relative 3D camera geometry. Let $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^3$ denote camera optical centers, $\mathbf{d}_i, \mathbf{d}_j \in \mathbb{S}^2$ their unit optical viewing directions, and $\mathcal{F}_i, \mathcal{F}_j \subset \mathbb{R}^3$ their bounded 3D camera frustums. We assess pairwise relationships using the Euclidean distance $r_{ij} = \|\mathbf{t}_i - \mathbf{t}_j\|_2$, the relative viewing angle $\theta_{ij} = \arccos(\mathbf{d}_i^\top \mathbf{d}_j)$,

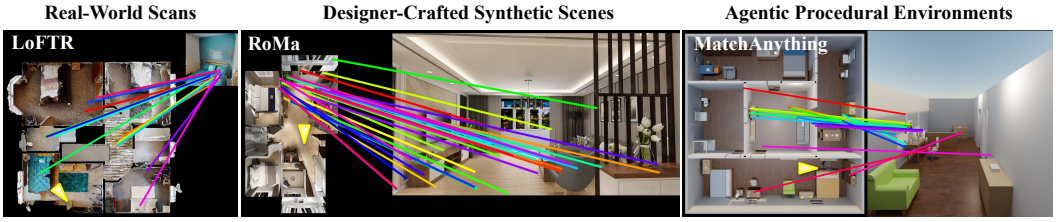

Figure 5: **Cross-view correspondence breakdown.** Off-the-shelf matchers (LoFTR (Sun et al., 2021), RoMa (Edstedt et al., 2023), MatchAnything (He et al., 2025)) fail to correlate in-situ captures with top-down views across diverse indoor environments. Yellow triangles on top-down views denote the ground-truth camera position and orientation, revealing that matches land far outside the true frustum. Due to disjoint surface visibility and drastic appearance changes across viewpoints, standard correspondence search fails.

and the signed ray closest-approach parameters $\mathbf{s} = [s_1, s_2]^\top$ along the optical axes. We build upon the viewpoint classification strategy in Xiangli et al. (2025) to categorize candidate pairs based on their baseline distance, mutual ray convergence, and frustum overlap. Pairs are assigned to the *high-overlap* regime if they share tight spatial baselines and forward-oriented, co-visible viewing cones. Conversely, pairs are assigned to the *zero-overlap* regime if they exhibit large spatial separation or proximate but disjoint configurations, such as head-on confrontation, divergent fields of view, or non-overlapping frustums. With pairwise overlap established, HIGH-OVERLAP is constructed greedily from high-overlap views, while ZERO-OVERLAP iteratively aggregates no-overlap views with every existing sequence item. Below is the detailed formulation and pair selection.

Ray Geometry and FoV Formulations. For a pair of views (I_i, I_j) with camera centers $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^3$ and unit viewing directions $\mathbf{d}_i, \mathbf{d}_j \in \mathbb{R}^3$, the closest approach parameters $\mathbf{s}^* = [s_1, s_2]^\top$ of their viewing rays are calculated via least-squares optimization:

$$\mathbf{s}^* = \arg\min_{\mathbf{s}\in\mathbb{R}^2} \left\| [-\mathbf{d}_i \quad \mathbf{d}_j]\, \mathbf{s} - (\mathbf{t}_i - \mathbf{t}_j) \right\|_2^2 . \tag{3}$$

The signs of s_1, s_2 indicate whether the closest-approach point lies in front of (> 0) or behind (< 0) the respective camera centers. To assess angular overlap across arbitrary camera aspect ratios, the diagonal field-of-view (FoV) Θ_{dia} is derived from the horizontal (Θ_h) and vertical (Θ_v) FoVs as:

$$\Theta_{\text{dia}} = 2 \arctan\left(\sqrt{\tan^2(\Theta_h/2) + \tan^2(\Theta_v/2)} \right). \tag{4}$$

Pair Selection Pipeline. Following Doppelgangers++ (Xiangli et al., 2025), frustum intersection $\mathcal{F}_i \cap \mathcal{F}_j \neq \varnothing$ is determined by a frustum intersection criterion bounded by the far depth $d_{\mathcal{F}}$. Algorithm 1 summarizes the rules for identifying HIGH-OVERLAP and ZERO-OVERLAP pairs. Threshold configurations across different scene subsets of TD-HOMES are detailed in Table 4.

Table 4: Threshold configurations for viewpoint sampling across scene subsets of TD-HOMES.

Subset	Sampling Threshold					
	$d_{\mathcal{F}}$	$r_{\mathcal{F}\text{pos}}$	r_{pos}	r_{neg}	θ_{opp}	θ_{front}
Real-World Scans	5 m	5 m	1.5 m	10 m	160°	80°
Designer-Crafted Synthetic	5 m	4 m	1.5 m	12 m	170°	80°
Agentic Procedural Environments	10 m	5 m	1.5 m	12 m	160°	80°

A.2.2 TRAINING DETAILS

To implement **HOME^3** (Decouple), the model is trained on 8 NVIDIA A100 GPUs for 80 epochs, with 800 iterations per epoch. During training, the DINO encoder remains frozen, and the model

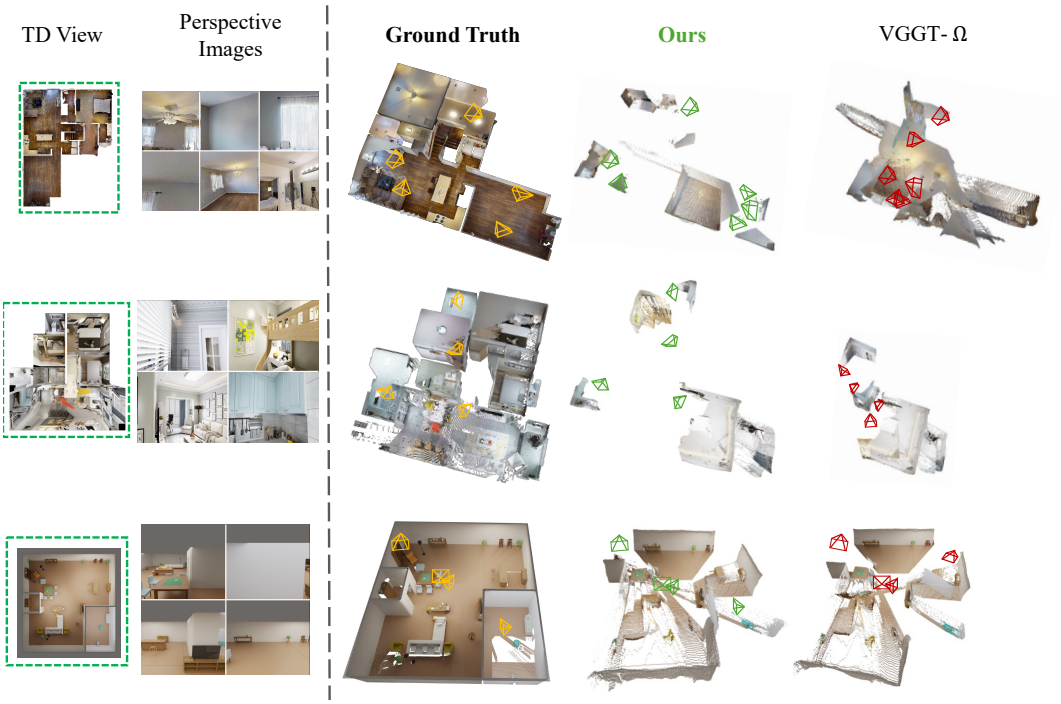

Figure 6: **Qualitative comparison of geometry estimation on multi-room scenes, implemented on VGGT-Ω.** We visualize both predicted camera poses (green/red cameras) and 3D pointmaps against the ground truth (yellow cameras). Existing pre-trained and fine-tuned baselines struggle to establish cross-room spatial relationships. In contrast, our **HOME^3** w/ *view-type decoupling* leverages the unposed top-down render to reliably estimate spatial relationships across non-overlapping perspective views, recovering more coherent global scene geometry.

is trained on images with random resolutions in the range of [100,000, 255,000], with aspect ratios ranging from 0.5 to 2.0. Following π^3, we implement the dynamic batch sizing strategy. Training samples are constructed sample-wise as image sequences with three levels of co-visibility: large overlap, small overlap, and no overlap. Each training sample comprises 3 to 24 images, containing exactly one top-down image and the remaining perspective images. Each GPU has a capacity of 24 images. For training objectives, we follow the implementations of π^3, concentrating on camera pose estimation and geometry reconstruction, therefore omitting the normal loss and confidence loss. We initialize the learning rate of all the trainable modules to be 5×10^{-6}. We adopt a `OneCycleLR` scheduler to decay the learning rate with a cosine annealing schedule.

A.3 Additional Experimental Results

More qualitative visualizations of **HOME^3** implemented on π^3 can be found in Fig. 7.

A.3.1 Generalization Analysis

Cross-Model Transferability. To assess the generalizability of **HOME^3** beyond π^3, we further transfer the proposed TD fusion strategy to VGGT-Ω. As shown in Tab. 5, the resulting **HOME^3** (Decouple) on VGGT-Ω improves 3D point map estimation across both in-domain and zero-shot benchmarks. These results indicate that **HOME^3** is not specific to a particular 3DFM backbone, but can be readily transferred to other advanced geometry models. Qualitative results are presented in Fig. 6.

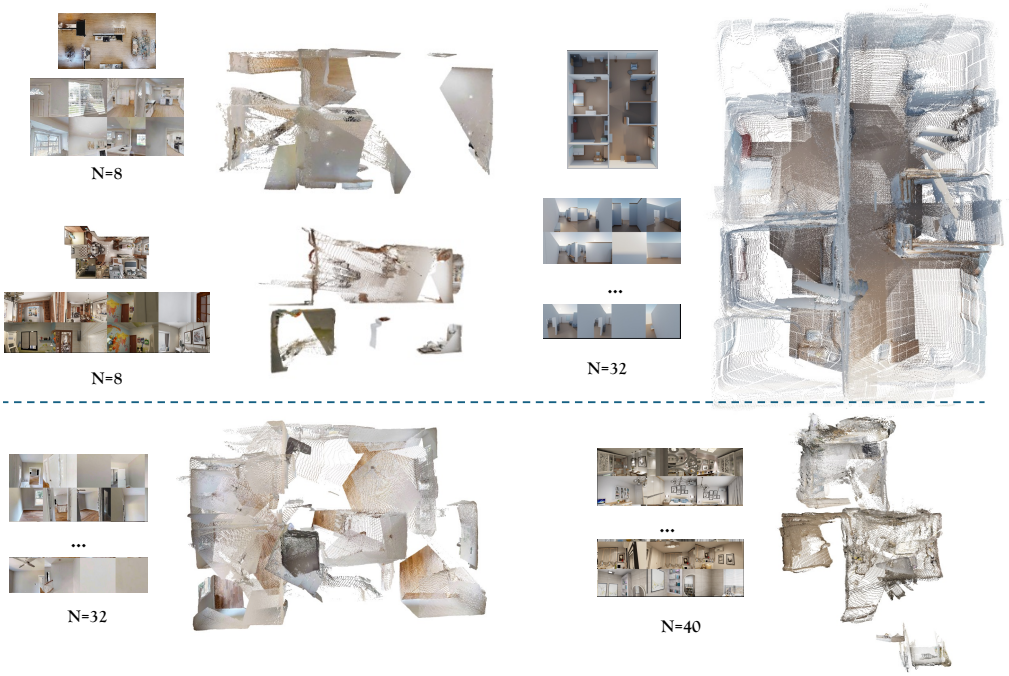

Figure 7: More visualization results. Top: results inferred with top-down view. Bottom: perspective-only estimation.

Table 5: **Cross-Model Transferability of our view-type decoupled design to VGGT-Ω, on Point Map Estimation.** We compare the pretrained VGGT-Ω using perspective views only with TD-VGGT-Ω using decoupled TD fusion. Best and second-best results within each benchmark are shown in bold and underlined, respectively.

Method	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT-Ω	<u>0.639</u>	<u>0.587</u>	<u>1.173</u>	<u>1.021</u>	<u>0.615</u>	<u>0.665</u>	<u>0.277</u>	<u>0.247</u>	<u>0.476</u>	<u>0.416</u>	<u>0.767</u>	<u>0.843</u>
HOME^3 (Decouple)	**0.302**	**0.246**	**0.367**	**0.277**	**0.752**	**0.873**	**0.152**	**0.122**	**0.154**	**0.117**	**0.840**	**0.958**

Perspective-only Generalization. We further evaluate whether our decoupled adaptation preserves zero-shot generalization to perspective-only settings, where top-down guidance is unavailable at inference. All methods are evaluated on ETH3D, Re10K, and NRGBD without top-down inputs.

As shown in Tab. 3, the Full Model largely retains the pre-trained 3DFM's performance and consistently outperforms **HOME^3** (Joint-FT) across these benchmarks. These results suggest that decoupling the attention projections (W_Q, W_K, W_V) and FFNs mitigates representational interference, enabling top-down-guided adaptation while preserving the pre-trained capability.

A.3.2 Ablation Studies

Ablation of View-Type Decoupling Design. Tab. 6 and Tab. 7 report the quantitative metrics of different view decoupling strategies applied to local and global attention layers. As shown in Tab. 6, decoupling global attention layers accounts for most of the improvement in point map estimation, with our design achieving the best overall reconstruction quality. For relative camera pose estimation (Tab. 7), decoupling both local and global attention layers yields marginally better results on the in-domain benchmark, whereas ours generalizes consistently better under the zero-shot setting.

Table 6: **Ablation study on the view-type decoupling design** for point map estimation. Local and Global briefly denote the local and global attention layers, respectively. We ablate the decoupling design on (a) neither of them, (b) both kinds of attention layers, and Ours.

Setting	Decouple		TD-HOMES						TD-HOMES (zero-shot)					
	Local	Global	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
			Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
(a)			0.295	0.240	0.362	0.277	0.765	0.889	0.194	0.156	0.218	0.173	0.801	0.928
(b)	✓	✓	0.280	0.226	0.334	0.258	**0.775**	**0.897**	0.181	0.148	0.207	0.170	0.818	0.944
Ours		✓	**0.278**	**0.223**	**0.333**	**0.252**	0.773	0.897	**0.161**	**0.132**	**0.178**	**0.148**	**0.826**	**0.948**

Table 7: **Ablation study on the view-type decoupling design** for relative camera pose estimation (angular). Local and Global briefly denote the local and global attention layers, respectively. We ablate the decoupling strategy on (a) neither of them, (b) both kinds of attention layers, and Ours.

Setting	Decouple		Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Local	Global	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES															
(a)			15.12	1.73	12.05	3.75	77.16	58.25	34.70	85.32	79.14	58.22	87.57	87.06	69.14
(b)	✓	✓	**13.84**	**1.57**	**11.18**	**3.22**	**79.64**	**61.98**	**38.60**	**86.97**	**80.87**	**61.36**	**88.81**	**88.12**	**71.67**
Ours		✓	14.02	1.59	11.32	3.30	79.33	61.53	38.04	86.74	80.82	60.96	88.70	87.96	71.46
TD-HOMES (zero-shot)															
(a)			3.47	1.15	9.90	2.04	93.35	75.63	50.90	97.45	86.63	72.33	98.20	89.25	80.06
(b)	✓	✓	2.75	1.22	9.58	1.88	95.45	77.84	52.54	98.26	87.92	74.42	98.69	89.61	81.52
Ours		✓	**2.37**	**0.91**	**9.05**	**1.53**	**96.21**	**80.97**	**59.68**	**98.67**	**88.32**	**77.50**	**99.05**	**89.83**	**83.28**

Ablation of Top-Down Supervision. We further investigate whether the top-down view should receive direct geometric supervision. Among the available objectives, only relative rotation supervision can be naturally defined for the top-down view. Since the top-down render is canonically aligned with the scene coordinate frame, its rotation can be represented by $\mathbf{I}_{3\times3}$, such that its relative rotation with each perspective view is directly determined by the ground-truth rotation of that view. In contrast, the top-down view does not provide a metrically aligned camera center or per-pixel depth, making direct supervision of relative translation and local point prediction ill-defined.

We therefore ablate whether the top-down view participates in the relative rotation objective. As shown in Tab. 9 and 8, excluding the top-down view from direct supervision consistently improves both relative camera pose estimation and point map reconstruction. These results support treating the top-down view as an exocentric condition rather than an additional reconstruction target.

Table 8: **Ablation study on top-down supervision** for point map estimation. We compare our default training strategy with a variant that additionally supervises the top-down view. Results are reported on both TD-HOMES and TD-HOMES (zero-shot). "w/ rot.-TD" denotes relative-rotation supervision on the TD view.

Setting	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
w/ rot.-TD	0.282	0.225	0.347	0.266	0.770	0.894	0.163	0.134	0.179	**0.144**	0.821	0.947
Ours	**0.278**	**0.223**	**0.333**	**0.252**	**0.773**	**0.897**	**0.161**	**0.132**	**0.178**	0.148	**0.826**	**0.948**

Table 9: **Ablation study on top-down supervision** for relative camera pose estimation. We compare our default training strategy with a variant that additionally supervises the top-down view. Results are reported on both TD-HOMES and TD-HOMES (zero-shot). "w/ rot.-TD" denotes relative-rotation supervision on the TD view.

Setting	Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES													
w/ rot.-TD	14.28	1.64	11.72	3.47	78.41	59.90	36.77	86.38	79.70	59.65	88.46	87.34	70.27
Ours	**14.02**	**1.59**	**11.32**	**3.30**	**79.33**	**61.53**	**38.04**	**86.74**	**80.82**	**60.96**	**88.70**	**87.96**	**71.46**
TD-HOMES (zero-shot)													
w/ rot.-TD	3.08	1.00	9.16	1.63	95.56	79.91	57.39	97.86	88.21	76.25	98.35	89.61	82.33
Ours	**2.37**	**0.91**	**9.05**	**1.53**	**96.21**	**80.97**	**59.68**	**98.67**	**88.32**	**77.50**	**99.05**	**89.83**	**83.28**

A.3.3 Experimental Results Across All Overlap Levels

We first supply the metrics of relative pose and multi-view depth estimation on TD-HOMES (Tab. 10, 11, and 12). These results provide an overall assessment of geometric robustness under heterogeneous viewing conditions. The quantitative results further show that our method delivers consistent improvements across both camera motion recovery and dense geometry estimation, indicating that the proposed design generalizes beyond a single geometric task.

Table 10: **Camera Pose Estimation on TD-HOMES**. Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) report the percentage of predictions within angular error thresholds of $5°$, $15°$, and $30°$. Rotation and translation errors are additionally reported by their mean and median values. Ours outperforms previous advanced methods.

Method	Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES													
ExtGeo (Zhang et al., 2025)	54.70	16.84	32.00	24.66	37.53	23.23	13.10	49.09	40.12	23.19	54.14	54.77	30.19
π^3 (Wang et al., 2026b)	53.24	11.15	31.24	23.56	41.75	25.31	15.57	52.04	41.16	25.33	56.06	55.99	32.09
π^3 (Perspective-FT)	34.62	3.43	23.96	12.36	55.83	34.44	21.18	65.12	53.70	35.16	69.75	68.09	44.65
HOME^3 (Joint-FT)	15.12	1.73	12.05	3.75	77.16	58.25	34.70	85.32	79.14	58.22	87.57	87.06	69.14
HOME^3 (Decouple)	**14.02**	**1.59**	**11.32**	**3.30**	**79.33**	**61.53**	**38.04**	**86.74**	**80.82**	**60.96**	**88.70**	**87.96**	**71.46**
TD-HOMES (zero-shot)													
ExtGeo (Zhang et al., 2025)	5.49	0.80	12.70	1.43	90.71	75.27	59.26	95.06	81.13	72.09	96.03	83.86	76.91
π^3 (Wang et al., 2026b)	4.54	**0.49**	12.03	**0.82**	93.21	77.59	**69.13**	95.90	81.36	76.09	96.91	84.56	79.35
π^3 (Perspective-FT)	4.79	1.41	13.16	2.60	91.95	68.76	42.77	97.07	81.06	65.92	97.32	84.65	74.17
HOME^3 (Joint-FT)	3.47	1.15	9.90	2.04	93.35	75.63	50.90	97.45	86.63	72.33	98.20	89.25	80.06
HOME^3 (Decouple)	**2.37**	0.91	**9.05**	1.53	**96.21**	**80.97**	59.68	**98.67**	**88.32**	**77.50**	**99.05**	**89.83**	**83.28**

A.3.4 Detailed Results across Data Sources and Overlap Levels

We report detailed experimental results on TD-HOMES from Tab. 13 to Tab. 24, covering HIGH-OVERLAP, ARBITRARY-OVERLAP, and ZERO-OVERLAP sequences. While the number of valid camera pairs varies with sequence length and overlap, each overlap level contains the same number of evaluation sequences. Accordingly, pairwise relative-pose metrics are computed by pooling all valid pose pairs within each test block, whereas trajectory, depth, and point-map metrics follow the task-specific aggregation protocols described in the corresponding table captions. The all-overlap results are recomputed from the union of the evaluation samples rather than obtained by taking an unweighted average of the three overlap-level results. Our method consistently improves geometric estimation, with pronounced gains as cross-view overlap decreases.

Table 11: **Relative Pose Estimation (Distance) on TD-HOMES.** High-Overlap, Arbitrary-Overlap, and Zero-Overlap sequences are jointly evaluated. TD-HOMES (zero-shot) denotes the agentic procedural environments used for zero-shot evaluation. Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) measure trajectory-level pose accuracy, where lower values are better. Ours consistently outperforms previous methods across both in-domain and zero-shot evaluations.

Method	TD-HOMES			TD-HOMES (zero-shot)		
	ATE ↓	RPE trans. ↓	RPE rot. ↓	ATE ↓	RPE trans. ↓	RPE rot. ↓
π^3 (Wang et al., 2026b)	2.377	4.494	69.75	0.932	1.948	19.11
ExtGeo (Zhang et al., 2025)	2.437	4.499	69.29	0.977	2.021	18.36
π^3 (Perspective-FT)	1.809	3.389	45.38	0.973	1.894	19.46
HOME^3 (Joint-FT)	0.887	1.809	21.23	0.445	0.912	10.61
HOME^3 (Decouple)	**0.829**	**1.680**	**19.26**	**0.364**	**0.788**	**8.34**

Table 12: **Multi-View Depth Estimation on TD-HOMES.** High-Overlap, Arbitrary-Overlap, and Zero-Overlap sequences are jointly evaluated. Absolute Relative Error (Abs Rel↓) and $\delta < 1.25 \uparrow$ measure depth estimation accuracy. Our method achieves strong performance across both in-domain and zero-shot evaluations.

Method	TD-HOMES		TD-HOMES (zero-shot)	
	Abs Rel ↓	$\delta < 1.25 \uparrow$	Abs Rel ↓	$\delta < 1.25 \uparrow$
π^3 (Wang et al., 2026b)	0.206	0.900	**0.051**	0.971
ExtGeo (Zhang et al., 2025)	0.204	0.900	0.052	0.970
π^3 (Perspective-FT)	0.076	0.946	0.058	0.967
HOME^3 (Joint-FT)	**0.075**	0.946	0.055	0.971
HOME^3 (Decouple)	0.076	**0.948**	0.052	**0.971**

Table 13: **Zero-Overlap Camera Pose Estimation** under pairwise relative-pose metrics. Statistics are computed after pooling all 4,509 TD-HOMES pose pairs, or all 280 TD-HOMES (zero-shot) pose pairs, within the corresponding test block. Lower errors and higher accuracy values are better.

Method	Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES													
π^3 (Wang et al., 2026b)	81.82	85.52	45.56	44.97	17.03	3.84	0.73	28.43	14.70	2.90	33.56	32.65	6.31
ExtGeo (Zhang et al., 2025)	81.43	84.53	46.07	45.63	13.28	3.61	0.52	26.28	14.79	2.34	32.31	31.83	5.53
π^3 (Perspective-FT)	59.15	40.59	38.91	35.12	25.35	8.34	2.87	37.77	24.17	8.38	46.40	44.13	15.44
HOME^3 (Joint-FT)	32.96	4.56	23.62	11.13	51.98	33.56	15.60	65.03	55.00	33.72	70.37	69.22	44.33
HOME^3 (Decouple)	**30.82**	**3.96**	**21.46**	**9.11**	**55.25**	**37.72**	**18.70**	**69.71**	**59.13**	**37.51**	**74.07**	**72.63**	**48.50**
TD-HOMES (zero-shot)													
π^3 (Wang et al., 2026b)	41.91	4.85	44.06	44.22	51.07	12.14	7.14	61.43	24.29	13.17	67.50	39.64	19.70
ExtGeo (Zhang et al., 2025)	36.99	5.35	46.27	48.61	48.21	13.93	6.29	65.71	22.50	13.21	68.93	34.64	18.51
π^3 (Perspective-FT)	39.63	6.06	42.96	41.17	44.29	12.86	4.21	63.93	26.79	13.10	67.86	41.43	20.40
HOME^3 (Joint-FT)	22.32	4.13	32.59	17.40	55.00	17.86	7.57	76.07	46.07	24.12	81.79	59.64	37.48
HOME^3 (Decouple)	**20.16**	**3.41**	**30.52**	**13.52**	**59.64**	**27.14**	**11.86**	**78.93**	**51.43**	**31.02**	**85.00**	**60.71**	**42.77**

Table 14: **Zero-Overlap Camera Pose Estimation** under trajectory-level metrics. TD-HOMES is the 162 : 301 sequence-count-weighted average of the overlap-specific real-world scans and designer-synthetic scenes dataset summaries; undefined RPE values from 18 single-frame designer-synthetic scenes samples are omitted within the designer-synthetic scenes summary. TD-HOMES (zero-shot) is the 80-sequence TD-Mansion summary. Lower ATE and RPE are better.

Method	TD-HOMES			TD-HOMES (zero-shot)		
	ATE ↓	RPE trans. ↓	RPE rot. ↓	ATE ↓	RPE trans. ↓	RPE rot. ↓
π^3 (Wang et al., 2026b)	2.893	6.322	95.03	1.935	4.145	51.61
ExtGeo (Zhang et al., 2025)	2.918	6.340	92.01	1.886	4.183	45.79
π^3 (Perspective-FT)	2.194	4.740	59.97	1.922	3.852	47.26
HOME^3 (Joint-FT)	<u>1.194</u>	<u>2.655</u>	<u>30.81</u>	<u>0.910</u>	**1.867**	<u>23.82</u>
HOME^3 (Decouple)	**1.075**	**2.408**	**27.57**	**0.863**	<u>1.958</u>	**22.38**

Table 15: **Zero-Overlap Multi-View Depth Estimation**. Within each dataset, metrics are weighted by the number of valid depth pixels; TD-HOMES then combines the real-world scans and designer-synthetic scenes summaries with sequence-count weights $162:301$. TD-HOMES (zero-shot) reports TD-Mansion. Lower Abs Rel and higher $\delta < 1.25$ are better.

Method	TD-HOMES		TD-HOMES (zero-shot)	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
π^3 (Wang et al., 2026b)	0.189	0.868	0.135	0.896
ExtGeo (Zhang et al., 2025)	0.187	0.870	<u>0.129</u>	0.893
π^3 (Perspective-FT)	0.084	0.927	0.135	0.902
HOME^3 (Joint-FT)	<u>0.080</u>	<u>0.931</u>	0.132	**0.906**
HOME^3 (Decouple)	**0.079**	**0.933**	**0.124**	<u>0.906</u>

Table 16: **Zero-Overlap Point Map Estimation**. Metrics are averaged over the 463 TD-HOMES sequences (162 real-world scans and 301 designer-synthetic scenes) or the 80 TD-HOMES (zero-shot) sequences from TD-Mansion. Lower accuracy and completeness errors and higher normal consistency are better.

Method	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
π^3 (Wang et al., 2026b)	0.970	0.918	1.677	1.548	0.584	0.625	0.641	0.602	0.958	0.845	0.641	0.705
ExtGeo (Zhang et al., 2025)	0.951	0.906	1.626	1.509	0.584	0.627	0.658	0.626	0.970	0.861	0.637	0.692
π^3 (Perspective-FT)	0.818	0.737	1.128	0.996	0.634	0.709	0.598	0.541	0.941	0.806	0.638	0.702
HOME^3 (Joint-FT)	<u>0.417</u>	<u>0.356</u>	<u>0.547</u>	<u>0.449</u>	<u>0.744</u>	<u>0.855</u>	<u>0.415</u>	<u>0.369</u>	<u>0.492</u>	<u>0.421</u>	<u>0.752</u>	<u>0.842</u>
HOME^3 (Decouple)	**0.379**	**0.319**	**0.497**	**0.400**	**0.756**	**0.869**	**0.369**	**0.333**	**0.445**	**0.392**	**0.788**	**0.878**

Table 17: **Arbitrary-Overlap Camera Pose Estimation** under pairwise relative-pose metrics. Statistics are computed after pooling all 12,964 TD-HOMES pose pairs, or all 2,078 TD-HOMES (zero-shot) pose pairs, within the corresponding test block. Lower errors and higher accuracy values are better.

Method	Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES													
π^3 (Wang et al., 2026b)	66.00	48.51	36.10	32.30	29.13	17.82	9.46	40.65	31.57	16.32	45.38	47.67	21.67
ExtGeo (Zhang et al., 2025)	67.54	57.20	37.21	34.39	25.34	15.94	7.73	37.31	30.34	14.49	43.31	45.81	19.70
π^3 (Perspective-FT)	44.95	8.85	27.84	18.65	43.62	27.04	14.90	54.77	45.36	26.37	60.86	61.52	35.30
HOME^3 (Joint-FT)	18.29	2.10	12.32	3.45	72.86	60.14	34.99	82.51	78.66	57.80	85.18	86.25	68.14
HOME^3 (Decouple)	**17.11**	**1.87**	**11.73**	**2.96**	**75.41**	**64.09**	**39.13**	**83.89**	**80.04**	**61.03**	**86.05**	**86.93**	**70.69**
TD-HOMES (zero-shot)													
π^3 (Wang et al., 2026b)	3.40	**0.60**	10.40	**0.76**	93.55	80.08	**70.65**	96.44	84.22	78.41	97.79	86.77	81.89
ExtGeo (Zhang et al., 2025)	5.33	0.93	11.17	1.28	90.47	77.09	60.44	95.48	83.93	73.97	96.63	86.04	79.24
π^3 (Perspective-FT)	3.04	1.52	11.12	2.16	93.74	73.48	45.68	98.89	84.22	69.66	98.89	87.34	77.67
HOME^3 (Joint-FT)	2.37	1.29	7.79	1.56	95.91	82.05	55.38	99.33	89.75	76.67	99.57	92.11	83.84
HOME^3 (Decouple)	**1.31**	1.00	**7.16**	1.24	**98.12**	**86.19**	64.91	**100.00**	**91.10**	**81.51**	**100.00**	**92.20**	**86.62**

Table 18: **Arbitrary-Overlap Camera Pose Estimation** under trajectory-level metrics. TD-HOMES is the $162:301$ sequence-count-weighted average of the overlap-specific real-world scans and designer-synthetic scenes dataset summaries; TD-HOMES (zero-shot) is the 80-sequence TD-Mansion summary. Lower ATE and RPE are better.

Method	**TD-HOMES**			**TD-HOMES (zero-shot)**		
	ATE ↓	RPE trans. ↓	RPE rot. ↓	ATE ↓	RPE trans. ↓	RPE rot. ↓
π^3 (Wang et al., 2026b)	3.341	6.117	84.96	0.563	1.086	4.82
ExtGeo (Zhang et al., 2025)	3.436	6.160	85.47	0.658	1.183	7.45
π^3 (Perspective-FT)	2.635	4.723	59.68	0.605	1.103	6.48
HOME^3 (Joint-FT)	1.255	2.484	26.35	0.256	0.545	4.62
HOME^3 (Decouple)	**1.199**	**2.345**	**24.45**	**0.133**	**0.255**	**1.56**

Table 19: **Arbitrary-Overlap Multi-View Depth Estimation**. Within each dataset, metrics are weighted by the number of valid depth pixels; TD-HOMES then combines the real-world scans and designer-synthetic scenes summaries with sequence-count weights $162:301$. TD-HOMES (zero-shot) reports TD-Mansion. Lower Abs Rel and higher $\delta < 1.25$ are better.

Method	**TD-HOMES**		**TD-HOMES (zero-shot)**	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
π^3 (Wang et al., 2026b)	0.205	0.895	**0.033**	**0.985**
ExtGeo (Zhang et al., 2025)	0.220	0.894	0.036	0.984
π^3 (Perspective-FT)	0.079	0.941	0.045	0.977
HOME^3 (Joint-FT)	**0.078**	0.943	0.041	0.982
HOME^3 (Decouple)	0.079	**0.943**	0.041	0.982

Table 20: **Arbitrary-Overlap Point Map Estimation**. Metrics are averaged over the 463 TD-HOMES sequences (162 real-world scans and 301 designer-synthetic scenes) or the 80 TD-HOMES (zero-shot) sequences from TD-Mansion. Lower accuracy and completeness errors and higher normal consistency are better.

Method	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
π^3 (Wang et al., 2026b)	0.742	0.673	1.503	1.259	0.565	0.597	0.126	0.088	0.199	0.150	0.812	0.921
ExtGeo (Zhang et al., 2025)	0.753	0.684	1.479	1.229	0.558	0.587	0.142	0.101	0.198	0.142	0.791	0.914
π^3 (Perspective-FT)	0.701	0.610	1.038	0.809	0.607	0.675	0.157	0.097	0.195	0.131	0.769	0.926
HOME^3 (Joint-FT)	0.384	0.311	0.449	0.326	0.739	0.869	0.095	0.054	0.092	0.054	0.818	0.970
HOME^3 (Decouple)	**0.376**	**0.298**	**0.417**	**0.303**	**0.749**	**0.876**	**0.063**	**0.035**	**0.052**	**0.029**	**0.841**	**0.984**

Table 21: **High-Overlap Camera Pose Estimation** under pairwise relative-pose metrics. Statistics are computed after pooling all 12,964 TD-HOMES pose pairs, or all 2,078 TD-HOMES (zero-shot) pose pairs, within the corresponding test block. Lower errors and higher accuracy values are better.

Method	Rot. Err. ↓		Trans. Err. ↓		@5°			@15°			@30°		
	Mean	Med.	Mean	Med.	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑	RRA ↑	RTA ↑	AUC ↑
TD-HOMES													
π^3 (Wang et al., 2026b)	30.53	2.27	21.39	8.68	62.96	40.28	26.84	71.64	59.96	42.14	74.58	72.44	51.48
ExtGeo (Zhang et al., 2025)	32.58	2.85	21.90	9.44	58.16	37.36	22.83	68.81	58.71	39.15	72.55	71.71	49.25
π^3 (Perspective-FT)	15.76	1.33	14.88	4.76	78.64	50.90	33.83	84.98	72.32	53.25	86.76	83.01	64.17
HOME^3 (Joint-FT)	5.74	1.15	7.76	3.17	90.20	64.96	41.06	95.19	88.01	67.15	95.94	94.06	78.78
HOME^3 (Decouple)	**5.08**	**1.08**	**7.40**	**2.92**	**91.62**	**67.26**	**43.68**	**95.52**	**89.15**	**69.03**	**96.43**	**94.32**	**80.21**
TD-HOMES (zero-shot)													
π^3 (Wang et al., 2026b)	**0.65**	**0.35**	9.35	**0.74**	98.56	**83.93**	**75.97**	**100.00**	86.19	**82.25**	**100.00**	88.40	84.84
ExtGeo (Zhang et al., 2025)	1.40	0.59	9.69	1.36	96.68	81.71	65.21	98.60	86.24	78.15	99.09	88.31	82.45
π^3 (Perspective-FT)	1.85	1.16	11.18	2.60	96.58	71.56	45.05	99.71	85.23	69.29	99.71	87.78	77.91
HOME^3 (Joint-FT)	2.03	0.95	8.96	2.19	95.96	77.00	52.25	98.46	88.98	74.49	99.04	90.38	82.02
HOME^3 (Decouple)	1.02	0.73	**8.06**	1.62	**99.23**	83.01	60.90	**100.00**	**90.52**	79.74	**100.00**	**91.39**	**85.41**

Table 22: **High-Overlap Camera Pose Estimation** under trajectory-level metrics. TD-HOMES is the 162 : 301 sequence-count-weighted average of the overlap-specific real-world scans and designer-synthetic scenes dataset summaries; TD-HOMES (zero-shot) is the 80-sequence TD-Mansion summary. Lower ATE and RPE are better.

Method	TD-HOMES			TD-HOMES (zero-shot)		
	ATE ↓	RPE trans. ↓	RPE rot. ↓	ATE ↓	RPE trans. ↓	RPE rot. ↓
π^3 (Wang et al., 2026b)	0.896	1.104	30.25	0.299	0.613	**0.90**
ExtGeo (Zhang et al., 2025)	0.957	1.069	31.27	0.387	0.696	1.84
π^3 (Perspective-FT)	0.596	0.756	17.07	0.393	0.725	4.65
HOME^3 (Joint-FT)	**0.212**	0.320	6.90	0.170	0.323	3.38
HOME^3 (Decouple)	0.214	**0.315**	**6.08**	**0.095**	**0.151**	1.07

Table 23: **High-Overlap Multi-View Depth Estimation**. Within each dataset, metrics are weighted by the number of valid depth pixels; TD-HOMES then combines the real-world scans and designer-synthetic scenes summaries with sequence-count weights $162:301$. TD-HOMES (zero-shot) reports TD-Mansion. Lower Abs Rel and higher $\delta < 1.25$ are better.

Method	TD-HOMES		TD-HOMES (zero-shot)	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
π^3 (Wang et al., 2026b)	0.208	0.926	<u>0.030</u>	**0.990**
ExtGeo (Zhang et al., 2025)	0.191	0.926	0.033	<u>0.990</u>
π^3 (Perspective-FT)	**0.066**	<u>0.962</u>	0.036	0.987
HOME^3 (Joint-FT)	<u>0.067</u>	0.960	0.034	0.989
HOME^3 (Decouple)	0.069	**0.963**	**0.030**	0.989

Table 24: **High-Overlap Point Map Estimation**. Metrics are averaged over the 463 TD-HOMES sequences (162 real-world scans and 301 designer-synthetic scenes) or the 80 TD-HOMES (zero-shot) sequences from TD-Mansion. Lower accuracy and completeness errors and higher normal consistency are better.

Method	TD-HOMES						TD-HOMES (zero-shot)					
	Acc. ↓		Comp. ↓		NC. ↑		Acc. ↓		Comp. ↓		NC. ↑	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
π^3 (Wang et al., 2026b)	0.248	0.207	0.440	0.341	0.689	0.774	0.076	0.052	0.150	0.113	<u>0.847</u>	0.939
ExtGeo (Zhang et al., 2025)	0.255	0.213	0.432	0.327	0.680	0.762	0.089	0.059	0.155	0.112	0.819	0.931
π^3 (Perspective-FT)	0.182	0.144	0.249	0.178	0.746	0.861	0.111	0.071	0.155	0.109	0.799	0.929
HOME^3 (Joint-FT)	<u>0.083</u>	<u>0.055</u>	<u>0.090</u>	<u>0.057</u>	<u>0.813</u>	<u>0.944</u>	<u>0.072</u>	<u>0.045</u>	<u>0.069</u>	<u>0.044</u>	0.833	<u>0.971</u>
HOME^3 (Decouple)	**0.080**	**0.053**	**0.086**	**0.054**	**0.816**	**0.946**	**0.049**	**0.028**	**0.038**	**0.021**	**0.848**	**0.984**

REFERENCES

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Xingyi He He, Hao Yu, Sida Peng, Dongli Tan, Zehong Shen, Hujun Bao, and Xiaowei Zhou. Matchanything: Universal cross-modality image matching with large-scale pre-training. *ArXiv*, 2025.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. \pi^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, 2026.

Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.